

Solution 1: High-dimensional Gaussian Distributions

We will use that $\|X\|_2^2$ is in fact a sum of squared Gaussian random variables:

$$\|X\|_2^2 = \sum_{i=1}^p X_i^2. \quad (\text{S})$$

(a) Using (S) we derive directly

$$\mathbb{E}(\|X\|_2^2) = \mathbb{E}\left(\sum_{i=1}^p X_i^2\right) = \sum_{i=1}^p \mathbb{E}X_i^2 = \sum_{i=1}^p 1 = p, \quad (\text{E})$$

since $X_i \sim \mathcal{N}(0, 1)$ for each $i = 1, \dots, p$. Again using (S) we obtain

$$\begin{aligned} \text{Var}(\|X\|_2^2) &= \text{Var}\left(\sum_{i=1}^p X_i^2\right) && (\text{Using (S)}) \\ &= \sum_{i=1}^p \text{Var}(X_i^2) && (X_1, \dots, X_p \text{ are i.i.d.}) \\ &= \sum_{i=1}^p (\mathbb{E}(X_i^4) - \mathbb{E}(X_i^2)^2) && (\text{Var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 \text{ for any rv } Y) \\ &= \sum_{i=1}^p (3 - 1) && (\text{Using the hint: } \mathbb{E}_{Y \sim \mathcal{N}(0,1)}(Y^4) = 3) \\ &= 2p. \end{aligned}$$

(b) We write

$$\|X\|_2 - \sqrt{p} = \underbrace{\frac{\|X\|_2^2 - p}{2\sqrt{p}}}_{=(1)} - \underbrace{\frac{(\|X\|_2^2 - p)^2}{2\sqrt{p}(\|X\|_2 + \sqrt{p})^2}}_{=(2)}.$$

It holds that

$$\begin{aligned} \mathbb{E}[(1)] &= \mathbb{E}\left[\frac{\|X\|_2^2 - p}{2\sqrt{p}}\right] \\ &= \frac{1}{2\sqrt{p}} (\mathbb{E}[\|X\|_2^2] - p) && (\text{Linearity of expected value}) \\ &= \frac{1}{2\sqrt{p}} (p - p) && (\text{Using (E)}) \\ &= 0. \end{aligned}$$

On the one hand, it holds that $0 \leq (2)$, as all terms are non-negative and consequently $0 \leq \mathbb{E}[(2)]$. On the other hand, since $\|X\|_2 \geq 0$ we have

$$\begin{aligned} (2) &\leq \frac{(\|X\|_2^2 - p)^2}{2p^{3/2}} \\ \Rightarrow \mathbb{E}[(2)] &\leq \mathbb{E}\left[\frac{(\|X\|_2^2 - p)^2}{2p^{3/2}}\right] = \frac{\text{Var}(\|X\|_2^2)}{2p^{3/2}} = \frac{1}{\sqrt{p}}. \end{aligned}$$

Putting everything together:

$$|\mathbb{E}(\|X\|_2 - \sqrt{p})| = \underbrace{|\mathbb{E}[(1)] - \mathbb{E}[(2)]|}_{=0} = \mathbb{E}[(2)] \leq \frac{1}{\sqrt{p}}. \quad (\text{b})$$

(c) The variance can be bounded as follows:

$$\begin{aligned}
\text{Var}(\|X\|_2) &= \text{Var}(\|X\|_2 - \sqrt{p}) && \text{(Variance does not change by constant shifts)} \\
&\leq \mathbb{E}[(\|X\|_2 - \sqrt{p})^2] && \text{(For any random variable } Y \text{ it holds that } \text{Var}(Y) \leq \mathbb{E}(Y^2)) \\
&= \mathbb{E}[\|X\|_2^2 - 2\sqrt{p}\|X\|_2 + p] \\
&= \mathbb{E}[\|X\|_2^2] - 2\sqrt{p}\mathbb{E}[\|X\|_2] + p && \text{(Linearity of expected value)} \\
&= 2p - 2\sqrt{p}\mathbb{E}[\|X\|_2] && \text{(Using (E))} \\
&= 2p - 2\sqrt{p}\mathbb{E}[\|X\|_2 - \sqrt{p} + \sqrt{p}] && \text{(Someone told us that it is a good idea)} \\
&= -2\sqrt{p}\mathbb{E}[\|X\|_2 - \sqrt{p}] && \text{(Linearity of expected value)} \\
&\leq 2\sqrt{p} \frac{1}{\sqrt{p}} = 2. && \text{(Using (b))}
\end{aligned}$$

(d) Since $Z = \frac{X-X'}{\sqrt{2}} \sim \mathcal{N}(0, \mathbf{I})$, we obtain from (b) and (c) that

$$|\mathbb{E}(\|Z\|_2 - \sqrt{p})| \leq \sqrt{\frac{1}{p}}, \quad (\text{d1})$$

$$\text{Var}(\|Z\|_2) \leq 2. \quad (\text{d2})$$

But

$$\|Z\|_2 = \sqrt{\sum_{i=1}^p \left(\frac{X_i - X'_i}{\sqrt{2}}\right)^2} = \sqrt{\frac{1}{2} \sum_{i=1}^p (X_i - X'_i)^2} = \sqrt{\frac{1}{2}} \sqrt{\sum_{i=1}^p (X_i - X'_i)^2} = \sqrt{\frac{1}{2}} \|X - X'\|_2. \quad (\text{d3})$$

Thus, (d1) implies

$$\begin{aligned}
|\mathbb{E}(\|Z\|_2 - \sqrt{p})| &\leq \sqrt{\frac{1}{p}} \\
\Leftrightarrow \sqrt{2}|\mathbb{E}(\|Z\|_2 - \sqrt{p})| &\leq \sqrt{\frac{2}{p}} \\
\Leftrightarrow |\mathbb{E}(\underbrace{\sqrt{2}\|Z\|_2}_{=\|X-X'\|_2} - \sqrt{2p})| &\leq \sqrt{\frac{2}{p}}.
\end{aligned}$$

Moreover, (d2) implies

$$\begin{aligned}
&\text{Var}(\|Z\|_2) \leq 2 \\
\Leftrightarrow 2\text{Var}(\|Z\|_2) &\leq 2 \cdot 2 \\
\Leftrightarrow \text{Var}(\sqrt{2}\|Z\|_2) &\leq 4 && (\text{Var}(aY) = a^2\text{Var}(Y) \text{ for any rv } Y \text{ and constant } a) \\
\Leftrightarrow \text{Var}(\|X - X'\|_2) &\leq 4. && \text{(Using (d3))}
\end{aligned}$$

(e) As for any $x, x' \in \mathbb{R}^p$ it holds that

$$\langle x, x' \rangle = \frac{1}{2}(\|x\|_2^2 + \|x'\|_2^2 - \|x - x'\|_2^2).$$

we can infer that

$$\begin{aligned}
\mathbb{E}\langle X, X' \rangle &= \frac{1}{2} (\mathbb{E}\|X\|_2^2 + \mathbb{E}\|X'\|_2^2 - \mathbb{E}\|X - X'\|_2^2) \\
&= \frac{1}{2} (p + p - \mathbb{E}\|X - X'\|_2^2) && \text{(Using (E))} \\
&= \frac{1}{2} \left(p + p - 2\mathbb{E} \underbrace{\frac{1}{2}\|X - X'\|_2^2}_{=\|Z\|_2^2} \right) && \text{(Using (d3))} \\
&= \frac{1}{2} (p + p - 2p) = 0. && \text{(Using again (E))}
\end{aligned}$$

For the variance we obtain

$$\begin{aligned}
 \text{Var}(\langle X, X' \rangle) &= \text{Var} \left(\sum_{i=1}^p X_i X'_i \right) \\
 &= \sum_{i=1}^p \text{Var} (X_i X'_i) && \text{(Independence)} \\
 &= p \text{Var} (X_1 X'_1) && \text{(Identical distributions)} \\
 &= p (\mathbb{E}[X_1^2 (X'_1)^2] - \mathbb{E}[X_1 (X'_1)]^2) \\
 &= p (\mathbb{E}[X_1^2] \mathbb{E}[(X'_1)^2] - \mathbb{E}[X_1]^2 \mathbb{E}[(X'_1)]^2) && \text{(Independence)} \\
 &= p. && (\mathbb{E}[X_1^2] = \mathbb{E}[(X'_1)^2] = 1 \text{ and } \mathbb{E}[X_1] = \mathbb{E}[X'_1] = 0)
 \end{aligned}$$

```
(f) # load library to sample from multivariate normal distribution
library(mvtnorm)

# compute average euclidean length of a matrix x (rows = samples)
average_euclidean_length <- function(x){
  mean(apply(x,1,norm,type="2"))
}

# compute variance of euclidean lengths of a matrix x (rows = samples)
variance_euclidean_length <- function(x){
  var(apply(x,1,norm,type="2"))
}

# compute average euclidean distances between matrices x and x2 (rows = samples)
average_euclidean_distances <- function(x,x2){
  z = c()
  for (i in 1:nrow(x)){
    z = rbind(z,x[i,]-x2)
  }
  mean(apply(z,1,norm,type="2"))
}

# compute variance of euclidean distances between matrices x and x2
variance_euclidean_distances <- function(x,x2){
  z = c()
  for (i in 1:nrow(x)){
    z = rbind(z,x[i,]-x2)
  }
  var(apply(z,1,norm,type="2"))
}

# compute average inner products between matrices x and x2 (rows = samples)
average_inner_product <- function(x,x2){
  z = c()
  for (i in 1:nrow(x)){
    z = rbind(z,x2%*%x[i,])
  }
  mean(z)
}

# compute variance of inner products between matrices x and x2 (rows = samples)
variance_inner_product <- function(x,x2){
  z = c()
  for (i in 1:nrow(x)){
    z = rbind(z,x2%*%x[i,])
  }
}
```

```

    }
    var(z)
  }

set.seed(5)

p_range      <- 2^seq(0,10)

avg_eucl_length <- c()
var_eucl_length <- c()
avg_eucl_dist   <- c()
var_eucl_dist   <- c()
avg_inner_prod  <- c()
var_inner_prod  <- c()

for (p in p_range){

  x          <- rmvnorm(n=100, mean=rep(0,p), sigma=diag(p))
  x2         <- rmvnorm(n=100, mean=rep(0,p), sigma=diag(p))

  avg_eucl_length <- c(avg_eucl_length, average_euclidean_length(x))
  var_eucl_length <- c(var_eucl_length, variance_euclidean_length(x))
  avg_eucl_dist   <- c(avg_eucl_dist, average_euclidean_distances(x,x2))
  var_eucl_dist   <- c(var_eucl_dist, variance_euclidean_distances(x,x2))
  avg_inner_prod  <- c(avg_inner_prod, average_inner_product(x,x2))
  var_inner_prod  <- c(var_inner_prod, variance_inner_product(x,x2))

}

# compare the results visually
par(mfrow=c(2,3))

plot(p_range, avg_eucl_length, type="l", main="Average Euclidean Length", xlab="p", ylab="")
lines(p_range, sqrt(p_range), col=2, lty=2)

plot(p_range, avg_eucl_dist, type="l", main="Average Euclidean Distances", xlab="p", ylab="")
lines(p_range, sqrt(2*p_range), col=2, lty=2)

plot(p_range, avg_inner_prod, type="l", main="Average Inner Products", xlab="p", ylab="")
abline(h=0, col=2)

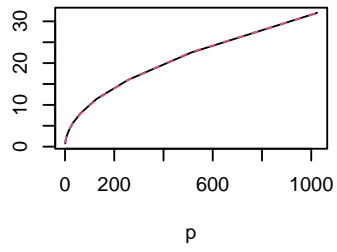
plot(p_range, var_eucl_length, type="l", main="Variance Euclidean Length", xlab="p", ylab="", ylim=c(0,2))
abline(h=2, col=2)

plot(p_range, var_eucl_dist, type="l", main="Variance Euclidean Distances", xlab="p", ylab="", ylim=c(0,4))
abline(h=2, col=2)

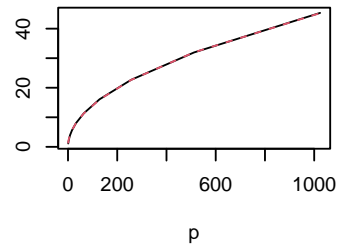
plot(p_range, var_inner_prod, type="l", main="Variance Inner Products", xlab="p", ylab="")
lines(p_range, p_range, col=2, lty=2)

```

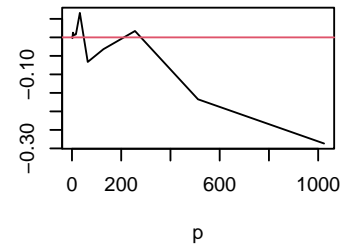
Average Euclidean Length



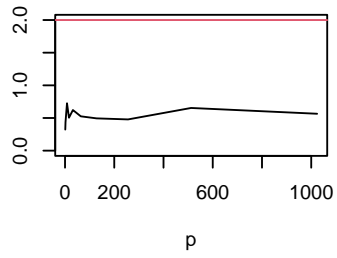
Average Euclidean Distances



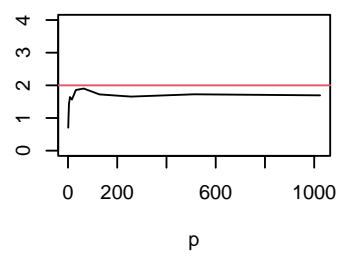
Average Inner Products



Variance Euclidean Length



Variance Euclidean Distances



Variance Inner Products

